Multi-task Joint Strategies of Self-supervised Representation Learning on Biomedical Networks for Drug Discovery

Xiaoqi Wang[1†], Yingjie Cheng[1†], Yaning Yang[1], Yue Yu[2], Fei Li[3*], and Shaoliang Peng[1,2,4*]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China.
[2]Peng Cheng Laboratory, Shenzhen, 518000, China.
[3]Computer Network Information Center, Chinese Academy of Sciences, Beijing 100850, China.
[4]The State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan University, Changsha 410082, China.

[†]These authors contributed equally to this work.
[*]To whom correspondence should be addressed; E-mail:
pittacus@gmail.com (Fei Li);
and slpeng@hnu.edu.cn (Shaoliang Peng).

Abstract

Self-supervised representation learning (SSL) on biomedical networks provides new opportunities for drug discovery. However, how to effectively combine multiple SSL models is still challenging and has been rarely explored. Therefore, we propose multi-task joint strategies of self-supervised representation learning on biomedical networks for drug discovery, named MSSL2drug. We design six basic SSL tasks inspired by various modality features including structures, semantics, and attributes in heterogeneous biomedical networks. Importantly, fifteen combinations of multiple tasks are evaluated by a graph attention-based multi-task adversarial learning framework in two drug discovery scenarios. The results suggest two important findings. (1) Combinations of multimodal tasks achieve the best performance compared to other multi-task joint models. (2) The local-global combination models yield higher performance than random two-task combinations when there are the same size of modalities. Therefore, we conjecture that the multimodal and local-global combination strategies can be treated as the guideline of multi-task SSL for drug discovery.

1 Introduction

Drug discovery is an important task for improving the quality of human life. However, it is an expensive, time-consuming, and complicated process that has a high chance of failure [1-2]. To improve the efficiency of drug discovery, a great number of researchers have devoted to developing or leveraging deep learning to speed up its intermediate steps, such as molecular property predictions [3-4], drug-target interaction (DTI)

predictions [5-11], and drug-drug interaction (DDI) predictions [12-13]. A key advantage behind these methods is that deep learning algorithms can capture the complex nonlinear relationships between input and output data [14].

In the past few years, deep learning techniques have gradually emerged as a powerful paradigm for drug discovery. Most deep learning architectures, such as convolutional neural networks [15] and recurrent neural networks [16], operate only on regular grid-like data (e.g., 2D images and text sequences), and are not well suited for graph data (e.g., DDI and DTI networks). However, in the real world, biomedical data are often formed as graphs or networks. In particular, biomedical heterogeneous networks (BioHNs) that integrate multiple types of data source are used extensively for life science researches. This is intuitive since BioHNs are well suited for modeling complex interactions in biological systems. For example, the BioHNs incorporating DDIs, DTIs, and protein-protein interactions (PPIs), protein-disease associations can naturally simulate the 'multi-drug, multi-target, multi-disease' biological processes within human body [17]. In the context of biomedical networks applications, graph neural networks (GNNs) [18-20], which are deep learning architectures specifically designed for graph structure data, are utilized to improve drug discovery. Such studies [21-24] use GNNs to generate the representation of each node in BioHNs, and formulate drug discovery as the node- or edge-level prediction problems. These graph neural network-based drug discovery approaches have shown high-precision predictions. However, most existing methods heavily depend on the size of training samples; that is, only large-scale training samples can help models to achieve great performance. Concurrently, with the variation of training sample sizes, the performance is changed by a large margin. Unfortunately, data labeling is expensive and time-consuming. Therefore, these graph-based deep learning models that rely on large-scale labeled data may not be satisfactory in real drug development scenarios.

Self-supervised representation learning (SSL) is a promising paradigm for solving the above issues. In SSL, deep learning models are trained via pretext tasks, in which supervision signals are automatically extracted from unlabeled data without the need for manual annotation. SSL aims to guide models to generate the generalized representations to achieve better performance on various downstream tasks. Following the immense success of SSL on computer vision [25-26] and natural language processing [27-28], SSL models built upon BioHNs have enjoyed increasing attention and have been successfully applied to drug discovery [29-32]. Unfortunately, most existing methods often design a single SSL task to train GNNs for drug discovery, thus leading to the built-in bias toward a single task and ignoring the multi-perspective characteristics of BioHNs. To cope with the potential bottleneck in single task-driven SSL applications, there are a few attempts leveraging multiple SSL tasks for facilitating performance of drug discovery [33-35]. These methods aim to integrate the advantages of various types of SSL tasks via the multi-task learning paradigms. However, most previous approaches train GNNs according to a fixed joint strategy involving multiple tasks, and do not focus on the differences between various multi-task combinations. Concurrently, the determination of which combination strategies can generate the most effective improvements has rarely been explored. Therefore, it is significant to pay

attention to the choice of multi-task combination strategies in SSL approaches. In addition, multi-task SSL methods built on BioHNs for drug discovery are still in the initial stages, and more systematic studies are pressingly needed.

To address the aforementioned problems, we propose multi-task joint strategies of self-supervised representation learning on biomedical networks for drug discovery, named MSSL2drug. Inspired by three modality features (i.e., structures, semantics, and attributes in BioHNs), six self-supervised tasks are developed to explore the impact of various SSL models on drug discovery. Next, fifteen multi-task joint strategies are evaluated via a graph attention-based multi-task adversarial learning model in two drug discovery scenarios. We find that the combinations of multimodal tasks can generate best performance compared to other multi-task strategies. Another interesting conclusion is that the local-global combination models tends to yield good results than random task combinations when there are the same sizes of modalities.

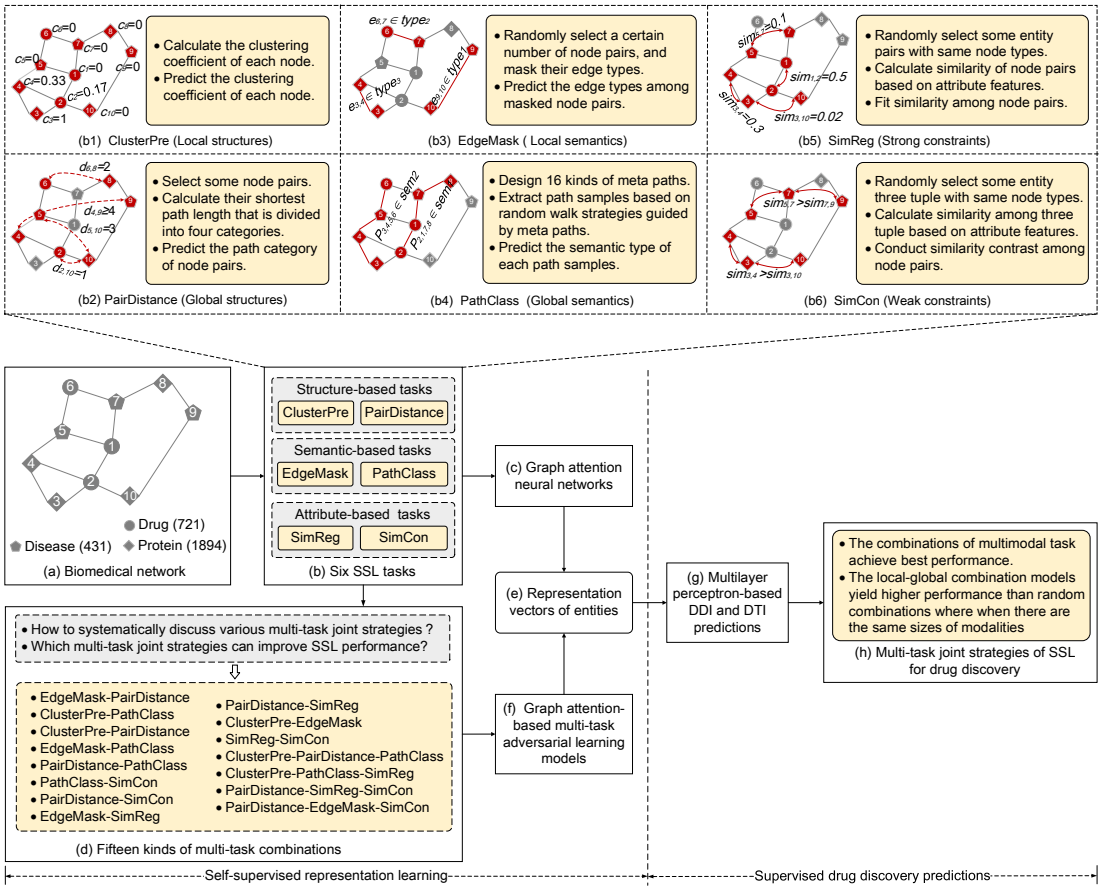

Figure 1. The schematic workflow of MSSL2drug. All circles, quadrangles, and pentagons denote the drugs, proteins, and diseases in a BioHN, respectively. The solid lines are the relationships among the biomedical entities in a BioHN. The red nodes represent the randomly selected vertices or node pairs in each of self-supervised task. The red solid lines in the edge type masked prediction (EdgeMask) and bio-path classification (PathClass) modules represent the randomly selected edges or paths. The dashed curved lines in the pairwise distance classification (PairDistance) module represent the measurements of the shortest paths between biomedical entities. The dashed solid lines in the node similarity regression (SimReg) and node similarity contrast (SimCon) modules represent the measurements of the similarities between biomedical entities. ClusterPre and PairDistance denotes clustering coefficient prediction and a pairwise distance classification, respectively.

2 Result

2.1 Overview of MSSL2drug

As shown in Fig. 1, we propose the multi-task joint strategies of self-supervised representation learning on biomedical networks for drug discovery, named MSSL2drug. First, we construct a biomedical heterogeneous network that integrates 3,046 biomedical entities and 111,776 relationships. Second, we develop six self-supervised tasks based on structures, semantics, and attributes in the BioHN, as shown in Fig. 1(b). These self-supervised tasks guide graph attention networks (GATs) to generate the representations from different views in the BioHN. More importantly, we develop fifteen kinds of multi-task combinations and a graph attention-based multi-task adversarial learning framework to improve the representation quality. Finally, the different representations from single-task and multi-task SSL are fed into multilayer perceptron (MLP) for predicting DDIs and DTIs. Based on the experiment results, we can draw two important findings. (1) The combinations of multimodal SSL tasks achieve state-of-the-art performance of drug discovery. (2) The joint training of local and global SSL tasks is superior to the random combinations of two SSL tasks when there are the same sizes of modalities.

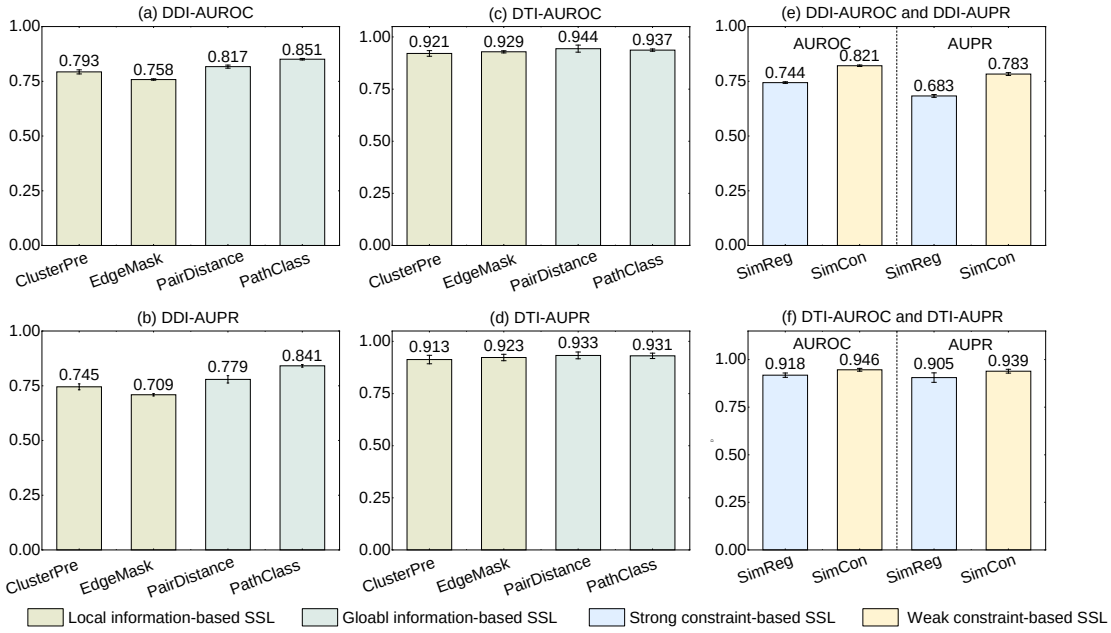

Figure 2. The results of single SSL tasks for drug warm start predictions, where the area under precision recall (AUPR) curve and area under receiver operating characteristic (AUROC) curve are used for the evaluation metrics. The error bars denote standard deviation values that are calculated across ten results.

2.2 Performance of single task-driven SSL on drug warm start predictions

In single task-driven SSL on drug warm start predictions, PairDistance and PathClass achieve relatively higher results, as shown in Fig. 2. Based on Student's t-test on the DTI and DDI results (as described in Supplementary Material Section S1), we find that PairDistance and PathClass significantly outperform ClusterPre and EdgeMask (*p*-

value <0.05). Another aspect to note is that SimCon is superior to SimReg. These results suggest that the global information-driven SSL approaches are superior to the local information-based SSL. Previous study [36] also made a similar finding. In addition, attribute weak constraint-based SSL tasks outperform strong constraint-based models.

2.3 Achieve superior performance by joint training the local and global SSL tasks in warm start scenarios

In this experiments, first, 11 two-task combination models are divided into two categories: single-modality combinations and two-modality combinations. It is noted that we design self-supervised tasks inspired by various modality knowledge including structures, semantics, and attributes in BioHNs. Therefore, there are up to three single-modality combination models, as shown in Fig. 3(b). Second, we compare the performance of two-task models with same size of modalities. The results in Fig. 3 suggest that the joint training of local and global SSL tasks (i.e., EdgeMask-PairDistance, ClusterPre-PathClass, ClusterPre-PairDistance and EdgeMask-PathClass) tends to obtain higher performance than random combinations of two SSL tasks when there are the same sizes of modalities. To further investigate the difference among various methods, we provide more analyses and a Student's t-test in Supplementary Material Section S2. Therefore, we conjecture that the local-global combination strategies can be regarded as an effective guideline for multi-task SSL to drug discovery.

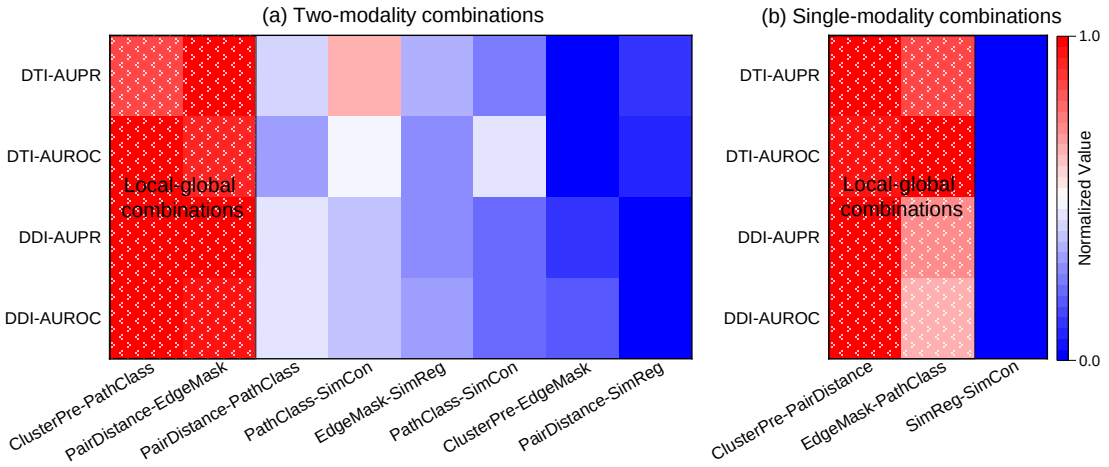

Figure 3. Heatmap of two-task combinations for drug warm start predictions where the results are normalized to [0,1] along the x-axis by the Min-Max normalization technique. The redder (bluer) squares denote the greater (smaller) the value. The shaded area denotes the combinations of global and local SSL tasks.

2.4 Achieve state-of-the-art performance by combining multimodal SSL tasks in warm start scenarios

The results in Fig. 4 show an interesting situation; that is, the growth of modalities leads to the significant performance improvement (*p*-value <0.05) for drug discovery. The more results and Student's t-test analyses can be found in Supplementary Material Section S3. These results suggest that combinations of multimodal tasks can achieve

best performance for drug discovery. Therefore, we conjecture that the multimodal combination strategy can be regarded as a potential guideline for multi-task SSL for drug discovery.

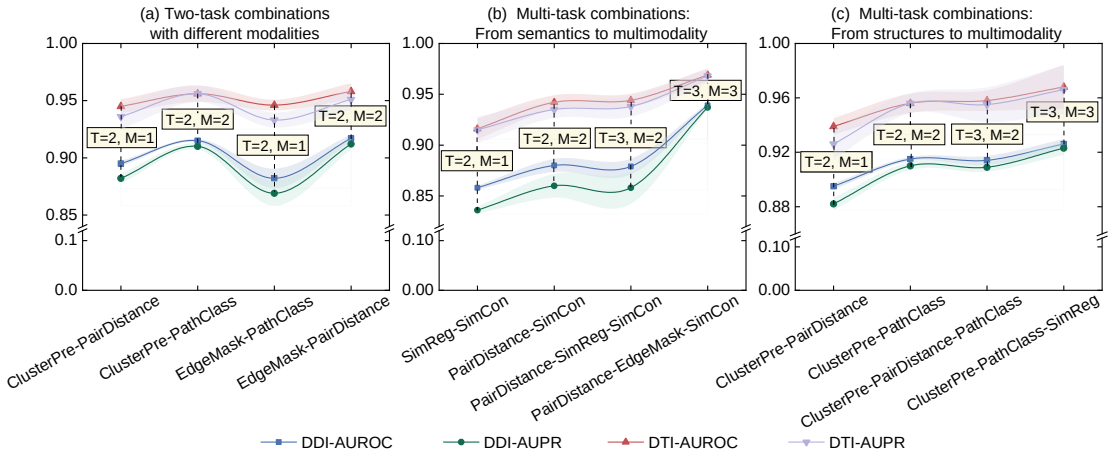

Figure 4. The results obtained multimodal task combinations for drug predictions, where 'T' and 'M' denote the total number of tasks and modalities in each multi-task combination, respectively. The shaded areas denote the standard deviation values that are calculated across ten results.

2.5 The results of cold start predictions demonstrate the influence of local-global and multimodal combination strategies on drug discovery

For the cold start drug prediction scenarios, the results of DDI and DTI predictions are generated by six basic SSL tasks and fifteen kinds of multi-task combinations. These results are straightforward and effective demonstrations that global information and attribute weak constraint-based SSL models can achieve better performance than local information and attribute strong constraint-based SSL. More importantly, these results verify that multimodal and local-global combination strategies can achieve state-of-the-art the prediction performance of drug discovery. The detailed result analyses can be found in Supplementary Material Section S4.

2.6 Performance validation of MSSL2drug on the external dataset

To demonstrate the robustness of MSSL2drug, it is used for Luo's dataset [6] and evaluated by warm and cold start predictions with different splitting ratios. The detailed setting and result analysis can be found in Supplementary Material Section S5. The results on warm start predictions suggest that the multimodal and local-global combination strategies still conducive to improving the performance of drug discovery on Luo's dataset. The performance of all SSL models on small training data and cold start predictions is reduced, because the volume of training set is reduced. However, we find the same performance distribution, that is, the multimodal and local-global combination strategies tend to generate better prediction performance. In addition, the results on different splitting ratios further demonstrate that MSSL2drug has the high robustness and generalization.

2.7 Performance comparisons

To demonstrate superiority of MSSL2drug, PairDistance-EdgeMask-SimCon are

compared with six state-of-the-art methods, including deepDTnet [37], MoleculeNet [38], KGE_NFM [5], DTINet [6], DDIMDL [39], and DeepR2cov [31]. On the constructed biomedical network data as shown in Table 1, we find that PairDistance-EdgeMask-SimCon is superior to six baselines. On the Luo's dataset as shown in Supplementary Material Table S12, PairDistance-EdgeMask-SimCon still outperforms other methods for warm start predictions. In addition, we compare MSSL2drug with Laplacian Eigenmaps (LE) [40], Graph Factorization (GF) [41] and DeepWalk [42], MF2A [43], and MIRACLE [44]. The results suggest that MSSL2drug can achieve higher performance on different datasets and scenarios. In addition, we compare MSSL2drug with Laplacian Eigenmaps (LE) [40], Graph Factorization (GF) [41] and DeepWalk [42], MF2A [43], and MIRACLE [44]. The results suggest that MSSL2drug can achieve higher performance on different datasets and scenarios. The more description can be found in Supplementary Material Section S6 and Section S16. We compare the run-time and parameter sizes in the Supplementary Material Section S12.

Table 1. Results of MSSL2drug and baselines for drug discovery predictions

Scenarios	Methods	DDI		DTI	
		AUROC±std	AUPR±std	AUROC±std	AUPR±std
Warm start	DeepR2cov	0.883±0.003	0.876±0.003	0.936±0.008	0.924±0.011
	DDIMDL	0.907±0.003	0.905±0.004	0.910±0.009	0.916±0.008
	DTINet	0.906±0.005	0.908±0.006	0.924±0.012	0.936±0.011
	KGE_NFM	0.914±0.004	0.915±0.004	0.923±**0.002**	0.935±**0.002**
	MoleculeNet	0.871±0.001	0.858±0.001	0.925±0.008	0.931±0.008
	deepDTnet	0.914±0.004	0.918±0.004	0.935±0.008	0.932±0.011
	PairDistance-EdgeMask-SimCon	**0.939**±**0.002**	**0.937**±**0.002**	**0.969**±0.006	**0.968**±0.007
Cold start	DeepR2cov	0.847±0.015	0.830±0.020	0.918±0.038	0.920±0.024
	DDIMDL	0.790±0.014	0.793±0.028	0.883±0.059	0.887±0.038
	DTINet	0.880±0.022	0.884±0.023	0.902±0.042	0.907±0.080
	KGE_NFM	0.734±0.018	0.722±0.034	0.906±**0.005**	0.886±**0.006**
	MoleculeNet	0.845±0.013	0.843±0.017	0.910±0.032	0.909±0.047
	deepDTnet	0.881±0.011	0.884±0.020	0.890±0.063	0.863±0.065
	PairDistance-EdgeMask-SimCon	**0.909**±**0.008**	0.895±**0.011**	**0.940**±0.020	**0.915**±0.048

a. 'std' denotes the standard deviation value calculated across ten results.

b. The best results are marked in **boldface**.

In addition, MSSL2drug and six baselines are evaluated under different splitting ratios between training sets and test sets, as shown in Supplementary Material Fig. S6. We observe that the performance of all methods are reduced when there are only few training samples. In particularly, when the ratio of training sets:test sets is 5:95 or 10:90, all methods achieve poor results for DDI and DTI predictions. An interesting finding is that the performance of MSSL2drug is without much fluctuation, and superior to baselines for different volume of training sets. These results suggest that most existing methods are prone to be influenced when applying to a small dataset, while MSSL2drug can partly overcome this limitation.

2.8 Case study: drug repositioning for COVID-19

Recently, the coronavirus disease 2019 (COVID-19) has posed a global health threat. Therefore, MSSL2drug is applied to drug repositioning for COVID-19, which aims to discover agents relating to IL-6 for blocking the excessive inflammatory response in patients. Based PubMed publications, clinical studies, molecular docking and molecular dynamics, we find that most of the predicted drugs may be able to inhibit the release of IL-6. More importantly, vandetanib (K_D=28.6μM) and pazopanib (K_D=20.7μM) can bind to IL-6 with high affinity as measured by SPR assay [45]. The detailed description are found in the Supplementary Material Section 7. However, it is necessary to further validate whether there are the indirect relationship or physical interactions between these drugs and IL-6 by standard and systematic experiments. In addition, all predicted drugs must be validated in preclinical experiments and randomized clinical trials before being used in patients.

2.9 Impact of key components on performance

2.9.1 Key component analyses in SSL tasks

- **Selection of centrality measurements in ClusterPre:** Compared to degree centrality, and eigenvector centrality [46], the clustering coefficient-based SSL model [47] achieves higher results (as described in Supplementary Material Section S8.1). A possible explanation for this result is that the clustering coefficients are not only extract the distribution of neighboring nodes, but also the triangle (loops of order 3) structures [48] in networks.
- **Division of "major" class in PairDistance:** The results suggest that dividing 4-hop and higher-hop node pairs into a "major" class achieves better performance compared to 3-hop and 5-hop (as described in Supplementary Material Section S8.2). This phenomenon is consistent with the finds in S^2GRL [49].
- **Length of meta path in PathClass:** We observe that selecting meta paths with lengths 4 is contribute to the performance of PathClass when compared to other length paths. Previous studies have made a similar finding [50-51]. The details can be found in Supplementary Material Section S8.3.
- **Selection of similarity measurement in SimCon:** We find that the different similarity measurements bring the marginal improvements or reductions to SimCon (as described in Supplementary Material Section S8.4). A possible explanation for this result is that SimCon only requires to distinguish the similarity distributions between node pairs, thus reducing the dependence on similarity measurements.
- **Ablation analyses of PairDistance-EdgeMask-SimCon:** We further suggest that integrating multimodal and local-global task is beneficial to improve performance of drug discovery. In PairDistance-EdgeMask-SimCon, the contribution of SimCon is relatively lower than EdgeMask and PairDistance to some extent (as described in Supplementary Section S8.5).

2.9.2 Key component analyses in multi-task learning framework

In this section, we evaluate the contribution of the adversarial training (ADL) strategy and orthogonality constraint (ORC) mechanism to MSSL2drug, respectively. The detailed description and results are available in Supplementary Material Section S9. We find that MSS2drug achieves best performance compared ADL and ORC models. In other words, MSS2drug integrating ADL and ORC is beneficial to improve performance of drug discovery. We also find that the contribution of ORC is higher than ADL to some extent. In addition, each task is trained by turn in a stochastic manner. The random task orders tend to be more robustness and reliability than the fixed task orders. However, we conjecture that using the prior or domain knowledge to set a specific order may contribute to the improvement of multi-task models.

2.9.3 High-quality Representation analyses

In this experiment, the representations from MSSL2drug are fed into Random forest (RF) [52] and support vector machine (SVM) [53] for drug discovery predictions, respectively. We find that using SVM and RF still achieves great performance for DDI and DTI predictions (as described in Supplementary Material Section S10). These results suggest that MSSL2drug can generate the high-quality representations that can keep the inherent nature of biomedical heterogeneous networks, thus improving the performance of drug discovery.

2.9.4 Dataset Contamination analyses

To understand how much influence does the data contamination in SSL have on DTI predictions, we remove the DTI of test set from self-supervised representation learning stage. The results suggest that the data contamination in SSL does not cause significant change for the performance of MSSL2drug. In other word, MSSL2drug is relatively insensitive to data contamination (as described in Supplementary Material Section S11).

3 Discussion

Recently, self-supervised representation learning on biomedical heterogeneous networks has emerged as a promising paradigm for drug discovery. Therefore, we aim to explore a combination strategy of multi-task self-supervised learning on biomedical heterogeneous networks for drug discovery. Based on six self-supervised learning tasks, we find that global knowledge-based SSL models outperform local information-based SSL models for drug discovery. This is intuitive and understandable since global view-based SSL tasks can capture the complex structures and semantics that is unable to be naturally learned by local SSL models. We also find that attribute weak constraint-based SSL tasks are superior to strong constraint-based models. This may be attributed to the fact that the similarity scoring functions are handcrafted and unable to accurately reflect the similarities among nodes in the original feature space. Unfortunately, the node similarity regression tasks arbitrarily fit node similarity values of node pairs. In contrast, similarity contrast tasks reduce the dependence on the original feature similarity values.

More importantly, fifteen kinds of multiple task combinations are evaluated by a graph attention-based multi-task adversarial learning model for drug discovery. These results suggest that the joint training the global and local tasks can achieve the relatively high prediction performance when there are the same sizes of modalities. In contrast, combining the tasks with great performance does not necessarily lead to better performance than other multi-task combinations for drug discovery. This is intuitive since there may be some conflicts and redundancies in the random combinations of SSL tasks. However, the combinations of global and local SSL models enable GNNs to leverage complementary information in BioHNs. To be specific, the local graph SSL models can capture the features within node itself or its first-order neighbors, but ignore the bird's-eye view of the node position in BioHNs. Fortunately, global SSL models can learn the dependencies among long-range neighborhoods, thus compensating the shortcomings of local SSL tasks. Simultaneously, an interesting finding is that combination models with multimodal tasks tend to generate best performance. This is because the combinations of multimodal tasks can capture multi-view information including structure, semantic and attribute features in BioHNs. The multimodal SSL models allow for knowledge transfer across multiple views and attain a deep understanding of natural phenomena in BioHNs. For a given SSL task, there are different levels of contributions in different multi-task combinations. Generally, if a SSL task can bring new modality information to multi-task models, it will generate the relatively greater contributions. In addition, if a local (global) information-driven SSL task is added to global (local) information-driven SSL tasks, it tends to bring high performance improvement. The multimodal and local-global combination strategies may be prioritized, when developing multi-task SSL for drug discovery. In other words, you can design yourself multi-task SSL models according to the multimodal and local-global combination strategies when you want to use MSSL2drug for drug discovery. On the other hand, you can also directly use PairDistance-EdgeMask-SimCon for drug discovery, because it integrates the multimodal and local-global SSL tasks, and achieves best performance.

In the application of deep learning, when there is a relative scarcity of labeled data, it is easy to cause the over-fitting problems, which exhibit a low testing performance even though its training performance is larger [54]. Fortunately, a great number of studies have suggested that multi-task learning techniques can greatly reduce the risk of over-fitting [55-59]. In particular, multi-task self-supervised learning can further overcome over-fitting issues and has emerged as a promising paradigm [54, 60-63]. The main reasons behind this are from two aspects. (1) SSL tasks drive deep learning models to learning the generalized representations from unlabeled data, thus reducing dependence on label data of downstream tasks (e.g., DDI predictions and DTI predictions). (2) Multi-task learning models can transfer and share knowledge among multiple SSL tasks to generate more general and informative representations. Therefore, multi-task self-supervised representation learning models like PairDistance-EdgeMask-SimCon can reduce the risk of over-fitting.

In conclusion, self-supervised representation learning based on biomedical heterogeneous networks provides new opportunities for drug discovery. To facilitate

this line of researches, we carefully explore the influence of various basic SSL tasks and propose unified combination strategies involving multi-task self-supervised representation learning to improve drug discovery. Simultaneously, we present a detailed empirical study to understand which combination strategies of multiple SSL tasks are most effective for drug discovery. In the future, we will pay attention to designing more SSL tasks and combination strategies to further improve performance of drug discovery. In addition, multi-class DDI predictions are closer to real-world drug discovery. Nevertheless, it is more challenging and difficult to manually annotate multi-class DDI data. Therefore, we will further verify the proposed global-local and multimodal combination strategies on multi-class DDI predictions.

4 Materials and methods

4.1 Biomedical heterogeneous networks

In this work, we construct a biomedical heterogeneous network (BioHN) according to deepDTnet [37]. The constructed BioHN assembles three types of nodes (i.e., drugs, proteins and diseases) and five types of edges (drug-drug interactions, drug-protein interactions, drug-disease associations, protein-protein interactions, and protein-disease associations). More specifically, the drug-drug interactions are extracted from the DrugBank database (v4.3) [64], where we only select drugs that have experimentally validated target information. The chemical name of each drug is transferred to a DrugBank ID. The drug-protein interaction networks are collected from the DrugBank database (v4.3), PharmGKB [65], and the Therapeutic Target database (TTD) [66]. We extract human protein-protein interactions with multiple pieces of evidences from the HPRD database (Release 9) [67], HuRI [68] and BioGRID [69]. Each protein name is transferred into an Entrez ID (https://www.ncbi.nlm.nih.gov/gene) via the NCBI (https://www.ncbi.nlm.nih.gov/). Drug-disease associations are attained via the fusion of the drug indications in the repoDB [70], DrugBank (v4.3), and DrugCentral databases [71]. Disease-protein associations are collected from two databases, including the Online Mendelian Inheritance in Man (OMIM) database [72] and the Comparative Toxicogenomics database (CTD) [73]. The disease names are standardized according to Unified Medical Language System (UMLS) vocabularies [74], and mapped to the MedGen ID (https://www.ncbi.nlm.nih.gov/medgen/) based NCBI database. BioHN in this work includes less information profiles than the dataset deepDTnet. Finally, the BioHN contains 3,046 nodes and 111,776 relationships (as described in Supplementary Material Table S26). There are 1,894 proteins, 721 drugs, and 4,978 drug-protein interactions in the BioHN. The ratio of DTI labels is $0.003 \approx 4{,}978/(721*1{,}894)$. Similarly, there are 66,384 drug-drug interactions in the BioHN. Therefore, the ratio of DDI label is $0.256 \approx 66{,}384/(721*720*0.5)$. In other words, there are sparse labels for DDI and DTI predictions. Therefore, we propose MSSL2drug that explore multi-task joint strategies of self-supervised representation learning on biomedical networks for drug discovery.

4.2 Basic self-supervised learning tasks

Multimodal information, including structures, semantics, and attributes in BioHNs,

provides unprecedented opportunities for designing advanced self-supervised pretext tasks. Hence, we develop six self-supervised tasks based upon the multimodal information contained in BioHNs for drug discovery.

4.2.1 Structure-based SSL tasks

The first direct choice for constructing SSL tasks is the inherent structure information contained in BioHNs. For a given node, self-supervision information is not only limited to itself or local neighbors, but also includes a bird's-eye view of the node positions in a BioHN. Therefore, we design a clustering coefficient prediction (ClusterPre) task that captures local structures and a pairwise distance classification (PairDistance) task that reflects the global structure information in BioHNs.

Clustering coefficient prediction (ClusterPre): In this pretext task, we use GATs to predict the clustering coefficient [47] of each node in BioHNs. The ClusterPre SSL task aims to guide GATs to generate low-dimensional representations that preserve the local structure information in BioHNs. In ClusterPre, the loss function adopts the mean squared error (as described in Supplementary Material Section S14.1).

Pairwise distance classification (PairDistance): We develop PairDistance that is not limited to a node itself and its local neighborhoods; it also takes global views of a BioHN. Similar to S^2GRL [49], we randomly select a certain number of node pairs and calculate the shortest path length between each node pair (i, j) as its distance value $d_{i,j}$. Subsequently, these node pairs and distance values are used to train GATs for drug discovery. In practice, the distances between node pairs are divided into four categories, that is, $d_{i,j}=1, d_{i,j}=2, d_{i,j}=3$ and $d_{i,j}\geq 4$. In other words, the PairDistance SSL task can be treated as a multiclass classification problem in which we adopt the cross entropy loss function (as described in the Supplementary Material Section S14.2). This is mainly attributed to two reasons. (1) The distinctions between the node pairs interacting via longer paths (i.e., $d_{i,j}\geq 4$) are relatively vague; thus, it is more reasonable to divide the longer pairwise distances into one "major" class [49]. (2) Based on the small-world phenomenon [75], we suppose that the shortest path lengths between most node pairs are within a certain range (as described in Supplementary Material Section S14.2). If we fit longer pairwise distances, some noisy values will be generated. Here, $d_{i,j}\geq 4$ indicates that PairDistance is not limited to the local connections in BioHNs. Therefore, PairDistance is beneficial for guiding GATs to generate node representation vectors that encode the global topology information of BioHNs. In addition, node pairs via random selection may lead to unstable results in PairDistance. Therefore, we repeat this process numerous times, and then the average performance is computed.

4.2.2 Semantic-based SSL tasks

BioHNs integrate multiple types of nodes or edges. The different relationships among these nodes contain distinct semantic information. Recent studies have suggested that semantic information can contribute to learning high-quality representations [28, 31]. Therefore, we develop edge type masked prediction (EdgeMask) task and bio-path classification (PathClass) task for encouraging GATs to capture certain aspects of semantic knowledge. Similar to the structure-based SSL tasks, EdgeMask and PathClass can capture the local and global semantics of BioHNs, respectively.

Edge type masked predictions (EdgeMask): This task is inspired by the BERT model [27], in which the core is a masked language model [76]. More specifically, we randomly mask edge types among some node pairs and then use GATs to predict these edge types, where the edge representation vectors are obtained by concatenating the representations of their two end-nodes. A detailed description of EdgeMask is found in Supplementary Material Section 14.3. The types of edges indicate the different action mechanisms between biomedical entities. Therefore, EdgeMask can enable GATs to learn the semantic features among local neighborhoods.

Bio-meta path classifications (PathClass): Compared to the types of edges among nodes, meta paths are a sequences for incorporating the complex semantic relationships in BioHNs (as described in the Supplementary Material Section S14.4). Different types of meta paths indicate distinct semantics. In PathClass, we design 16 types of meta paths as shown in the Supplementary Material Table S27, where the first or last objects are drugs or proteins, respectively. This is mainly because drugs and proteins are interconnected with other entities by more edges (as described in Supplementary Material Table S28). These meta paths guide random walks to extract path samples from BioHNs. In addition, we generate an equal number of false path instances by randomly replacing some nodes in true path instances. To be specific, for a given true path instance, it has 6.25% (i.e., 1/16) chance being replaced to generate a false path instance (as described in Supplementary Material Section S14.4). Therefore, all path samples are divided into 17 categories, including 16 kinds of true meta paths and one kind of false meta paths. Finally, we use GATs to predict the type of each path sample for learning node representations that contain rich semantics and complex relationships. Similarly, we adopt the cross entropy as loss function in PathClass.

4.2.3 Attribute-based SSL tasks

In addition to structures and semantics, attribute features play key roles in self-supervised representation learning. More generally, nodes with similar properties, such as the simplified molecular input line entry system (SMILES) strings [77] of drugs, should be distributed closely in the representation space. However, GATs only aggregate the features of node itself and its local neighborhoods, thus losing the similarity features among nodes. Based on this intuition, we develop two attribute-based SSL tasks, i.e., node similarity regression (SimReg) and node similarity contrast (SimCon), to enable GATs to maintain the similarity attributes in the original feature space. According to the degree of dependence on the original feature similarities, SimReg and SimCon can be categorized as strong constraint- and weak constraint-

based SSL paradigms, respectively.

Node similarity regression (SimReg): The proposed SimReg task requires GATs to fit similarity values of node pairs. More specifically, we randomly select a certain number of node pairs (i, j) (where i and j are the same types of nodes); and then calculate their similarity value $sim_{i,j}$ in the original feature space, such as the similarity between drug SMILES sequences. We require GATs to fit the similarity values ($sim_{i,j}$) of node pairs in the original feature space as possible. In other words, SimReg encourages GATs to learn representations via a strong constraint-based SSL paradigm. In this work, we use different property similarity measurements in the various types of nodes. The Tanimoto coefficient [78] among the SMILES sequences of drugs are treated as drug-drug similarity scores. We leverage the Smith-Waterman algorithm [79] to calculate the sequence similarity scores of protein pairs. The disease similarity scores are obtained by using PPI-based ModuleSim algorithm [80]. The detailed similarity measurement approaches and objective functions are described in Supplementary Material Section 14.5.

Node similarity contrast (SimCon): In SimReg, the similarity scoring mechanisms have an important impact on the representation learning process. SimReg cannot guarantee to generate the high-quality representations when similarity scores may not accurately reflect the true similarity values among nodes in original feature space. Therefore, we propose SimCon to reduce the influence of similarity scoring mechanisms. In SimCon, it assumes that the similar nodes in the original feature space should be closer in the embedding space than dissimilar nodes. More specifically, we randomly select a certain number of three tuples (i, j, k) for nodes, where i, j and k belong to the same types of nodes and $sim_{i,j} \geq sim_{i,k}$. For a given tuple (i, j, k), we use GATs to conduct a node similarity contrast; that is, the cosine values ($\cos_{i,j}$ and $\cos_{i,k}$) between the node representations generated by GATs should satisfy $\cos_{i,j} \geq \cos_{i,k}$. Formally, we propose a novel objective function:

$$\ell_{simCon}(\theta) = \frac{1}{|M|} \sum_{(i,\ j,k) \in M} L(i,\ j,k) \tag{1}$$

where M is the selected set of three tuples $(i,\ j,k)$, $|M|$ is the number of three tuple, and $L(i,\ j,k)$ is calculated as follows:

$$L(i,\ j,k)=\begin{cases}0, & \cos\left(\ f_\theta(i),\ f_\theta(j)\right)-\cos\left(\ f_\theta(i),\ f_\theta(k)\right)\geq 0\\ g(i,j,k), & \text{otherwise}\end{cases} \tag{2}$$

where $g(i,j,k)$ is calculated as follows:

$$g(i,j,k)=sim_{i,j}-sim_{i,k}-\left(\cos\left(\ f_\theta(i),\ f_\theta(j)\right)-\cos\left(\ f_\theta(i),\ f_\theta(k)\right)\right) \tag{3}$$

where θ is the parameters of a graph neural network $f_\theta(\cdot)$ and $f_\theta(i)$ denotes the embedding vectors of node i. In addition, $\cos(\cdot,\cdot)$ is the cosine similarity value between two embedding vectors.

Obviously, SimCon only requires that GATs can distinguish the similarity distributions between node pairs (i,j) and node pairs (i,k). However, SimReg requires that GATs fit similarity values for node pairs. Therefore, SimCon reduces the dependence on the original feature similarity values compared to SimReg; thus, SimCon is a weak constraint-based SSL paradigm.

4.3 Graph attention-based multi-task adversarial learning

In this work, the integration of the multi-task learning and GATs is a challenging and critical problem. Inspired by [81], we propose a graph attention-based adversarial multi-task learning framework for drug discovery, as shown in Fig. 5. The graph attention-based multi-task adversarial learning framework can be divided into the private and share parts that employ graph attention networks (GATs) [19] with different parameters.

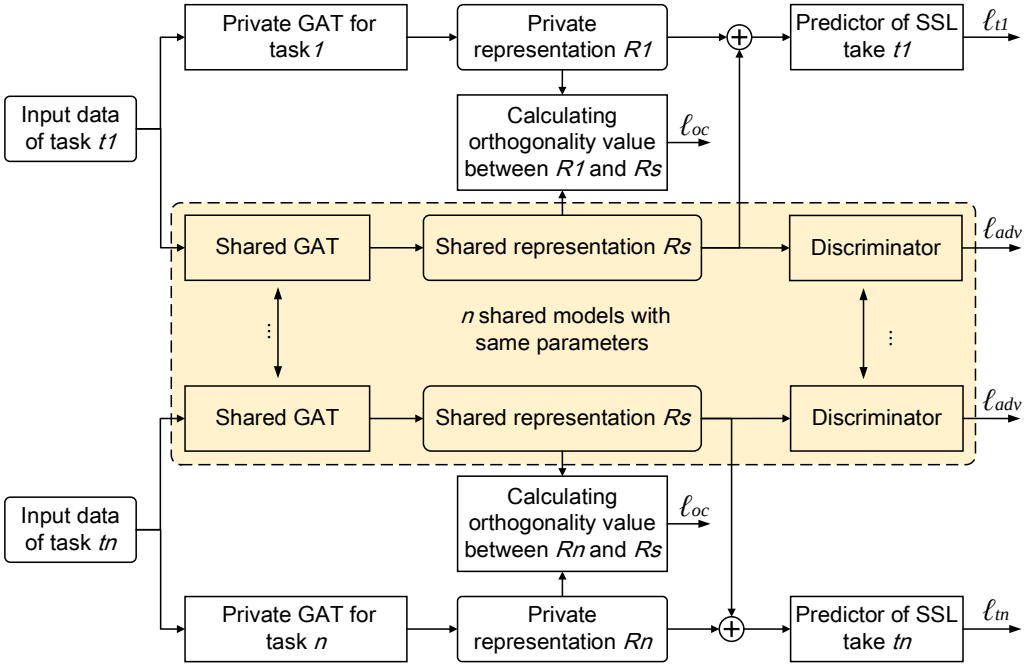

Figure 5. The framework of graph attention-based adversarial multi-task learning. For each epoch, we randomly select a SSL task *tn* from multi-task combinations. The corresponding private and shared GAT models generate the task-specific representations (*Rn*) and common representations

(*Rs*), respectively. The *Rn* and *Rs* are concatenated, and then fed into the MLP-based predictor of SSL task *tn*. In addition, the *Rs* are fed into the MLP-based discriminator to predict what kind of task the shared representation vectors come from. The parameters of current private and shared GAT models are updated by back-propagation based on the loss values from a SSL task predictor and discriminator, respectively. Finally, the parameters of the current shared model are assigned to all the other shared models. Therefore, we attain *n* private GAT models and shared GAT models with same parameters after multi-task SSL training. In other words, MSSL2drug generates the private representations by all private GATs and the shared representation by an arbitrary shared model.

4.3.1 Graph attention network

The graph attention network (GAT) is a popular graph neural network. GAT assumes that the contributions of neighboring nodes to the central nodes are different. To calculate the representations of one node, GAT aggregates its neighbor features by a multi-head attention mechanism. For a given node, the features from multiple attention mechanism models are concatenated to generate the final representation vectors. The final output features of each node can be calculated by:

$$\vec{h}_i' = \Big\Vert_{k=1}^{K} \sigma\left(\sum_{j \in N_i} \alpha_{ij}^k \mathbf{W}^k \vec{h}_j \right) \tag{4}$$

where σ is a nonlinearity activation function, K is the number independent attention mechanisms, $\mathbf{W}^k$ is the weight matrix of linear transformation in the *k*-th attention mechanism, N_i is the number of neighbors of node i, $\|$ represents concatenation operation, $\vec{h}_j$ is the current representations of neighbor j. More importantly, α_{ij}^k is the attention coefficients computed by the *k*-th attention mechanism. Intuitively, there are K attention coefficients between node i and j. α_{ij}^k can be calculated by:

$$\alpha_{ij}^k = \frac{\exp\left(\mathrm{LeakyReLU}\left((\vec{\mathrm{a}}^k)^{\mathrm{T}}\left[\mathbf{W}^k\vec{h}_i \,\middle\|\, \mathbf{W}^k\vec{h}_j\right]\right)\right)}{\sum_{j \in N_i} \exp\left(\mathrm{LeakyReLU}\left((\vec{\mathrm{a}}^k)^{\mathrm{T}}\left[\mathbf{W}^k\vec{h}_i \,\middle\|\, \mathbf{W}^k\vec{h}_j\right]\right)\right)} \tag{5}$$

where $\vec{\mathrm{a}}^k$ is a weight vector in *k*-th attention mechanism, $(\cdot)^{\mathrm{T}}$ represents transposition.

4.3.2 Task discriminator

For any node i in task t, the shared GAT generates task-invariant representations $x_t^i = f_{\theta_s}(i)$ where θ_s is the parameter of the shared GAT $f_{\theta_s}(\cdot)$. Then, these representation vectors x_t^i are fed into a multilayer perceptron that is treated as a task discriminator. This multilayer perceptron aims to predict what kind of task the shared representation vectors come from.

$$D(x_t^i, \theta_{td}) = \text{softmax}\left(\text{MLP}_{\theta_{td}}(x_t^i)\right) \tag{6}$$

where $\text{MLP}(\cdot)$ is a multilayer perceptron in which the trainable parameter is θ_{td}.

The loss values from the task discriminator can be calculated as follows:

$$\ell_{adv} = \min_{\theta_s}\left(\max_{\theta_{td}} \sum_{t=1}^{T}\sum_{i=1}^{N_t} y_t^i \log\left[D\left(f_{\theta_s}(i), \theta_{td}\right)\right]\right) \tag{7}$$

where N_t is the number of training nodes in task t, and y_t^i denotes the ground-truth labels indicating the type of current task.

4.3.3 Orthogonality constraints

The above shared model generates some features that may appear in both the shared space and the private space. Therefore, we adopt an orthogonality constraint [81-82] to eliminate redundant features from the private and shared spaces. Formally, the objective function of the orthogonality constraint is calculated as follows:

$$\ell_{oc} = \sum_{t=1}^{T}\sum_{i=1}^{N_t}\left\| f_{\theta_t}(i)^{\mathrm{T}} \cdot f_{\theta_s}(i)\right\|_F^2 \tag{8}$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm, and $f_{\theta_t}(\cdot)$ is the private GAT of the current task t.

4.3.4 Multi-task adversarial training

The final loss function of multi-task SSL can be written as follows:

$$\ell_{total} = \ell_t + \lambda \ell_{adv} + \gamma \ell_{oc} \tag{9}$$

where λ and γ are hyperparameters. ℓ_t denotes the loss value of task t.

During multi-task learning phase, inspired by [83], the models are trained in a stochastic manner by looping over the tasks.

***Step 1*:** Randomly select a task.

Step 2: Sample an epoch of instances from the task and train the corresponding private model and shared model.

Step 3: Update the corresponding parameters by back-propagation. Subsequently, the parameters of the current shared model are assigned to all the other shared models.

Step 4: Go to Step 1.

In this way, multiple private and shared GAT models are updated by the corresponding specific task. However, in practice, these shared GAT models are equivalent to a GAT model because they have the same parameters. In other words, we attain multiple private GAT models and a shared GAT model. Therefore, in two-task learning cases, Fig. S17(a) is equivalent to Fig. S17(b) in Supplementary Materials.

For a given node, different SSL tasks in different epochs guide the shared GAT to

capture the features with itself task property. Therefore, self-supervised training in different epochs can be treated as the adversarial learning process, that is, each SSL task encourages shared GAT to generate task-specific representations. After sufficient training, the shared GATs reach a point, at which it integrates the property of different tasks. Therefore, the shared feature space simply contains common information. In contrast, the private GAT model generates task-specific representations to make accurate SSL predictions.

4.4 Initialization features

In MSSL2drug, the initialization features of each node and adjacency matrixes of BioHNs are fed into GATs to perform training and test. Here, we take an example to describe the process of feature initialization, as shown in Supplementary Material Fig. S18. There are three key steps to generate the initialization features. For each given node, its neighbors are divided into three categories (i.e., drugs, proteins, and diseases).

***Step 1*:** Counting the number of neighbors in each class, $X=\{x_1, x_2, \cdots, x_N\}$, $Y=\{y_1, y_2, \cdots, y_N\}$ and $Z=\{z_1, z_2, \cdots, z_N\}$, respectively, where N is the total number of nodes. For instance, for given *node 1*, x_1=1, y_1=2, z_1=1, the sum of x_1, y_1, z_1 is its degree (i.e., the number of its neighbors), as shown 1st row in Supplementary Material Fig. S18(b);

***Step 2*:** Converting X, Y and Z to matrixes $\mathbf{X}=\{\vec{u}_1, \vec{u}_2, \cdots, \vec{u}_N\}$, $\mathbf{Y}=\{\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_N\}$, and $\mathbf{Z}=\{\vec{g}_1, \vec{g}_2, \cdots, \vec{g}_N\}$ by one-hot encoding technologies (https://www.educative.io/blog/one-hot-encoding);

***Step 3*:** Generating initialization feature matrix $\mathbf{F}=\{\vec{u}_1 \| \vec{v}_1 \| \vec{g}_1, \vec{u}_2 \| \vec{v}_2 \| \vec{g}_2, \cdots, \vec{u}_N \| \vec{v}_N \| \vec{g}_N\}$ by concatenating $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, where $\|$ is a concatenation operation.

4.5. Experiment settings

4.5.1 Multi-task combination settings

We design various multi-task combinations to answer two key questions.

- **Can joint training of two tasks with great performance (like 'Alliance between Giants') achieve higher performance than random combination of two tasks?** The results of single task SSL suggest that PairDistance, PathClass, and SimCon achieve the relatively higher performance. Therefore, we first chose all combinations of 'Alliance between Giants' (i.e., PairDistance-PathClass, PathClass-SimCon, and PairDistance-SimCon). Next, we randomly select 8 other two-task combinations, i.e., EdgeMask-PairDistance, ClusterPre-PathClass, ClusterPre-PairDistance, EdgeMask-PathClass, EdgeMask-SimReg, PairDistance-

SimReg, ClusterPre-EdgeMask, SimReg-SimCon.

- **Can the combinations integrating multimodal information further improve the prediction performance?**

  Based on 11 two-task combinations, we select four multi-task combinations to evaluate the influence of different modalities. As shown in Table S29 in Supplementary Materials, there is only one different task in context compositions. For example, PairDistance-SimCon is turned into PairDistance-EdgeMask-SimCon by adding SimReg. In addition, the pool of combination strategies keep diversity criterions, that is, each task is combined at least 5 times. Therefore, we selects 15 kinds of task combinations to guaranteed reliability.

4.5.2 Drug discovery predictions under different scenarios

In this study, we focus on the performance of various SSL tasks on DDI and DTI predictions, because they are key stages and play important roles in various applications of drug discovery. Simultaneously, DDI and DTI predictions are treated as link predictions in homogeneous and heterogeneous networks, respectively. Therefore, DDI and DTI predictions can systematically demonstrate the performance of various kinds of SSL tasks and combination strategies. According to the guidance of KGE_NFM [5], we design the following two experimental scenarios. **Warm start predictions:** Given a set of drugs and their known DTIs, we aim to predict other potential interactions between these drugs. All the known interactions are positive samples, and an equal number of negative samples are randomly selected from the unknown interactions. The positive and negative samples are split into a training set (90%) and a testing set (10%). In this situation, the training set may include drugs and targets contained in the test set. The same experimental setting as DTI predictions are used for DDI predictions. In this experimental scenarios, we compare the differences among various SSL tasks for DDI and DTI predictions, and draw a conclusion on which combination strategies can generate the best performance. **Cold start for drugs:** In real drug discovery, it is more important and challenging to predict potential targets and drugs that may interact with newly discovered chemical compounds. In other words, the test set contains drugs that are unseen in the training set. To be specific, we randomly select 5% drugs, and then all DTI and DDI pairs associated with these drugs are treated as test set. This scenario aims to validate the conclusions that are found in the warm start predictions. We use the area under precision recall (AUPR) curve and area under receiver operating characteristic (AUROC) curve as the evaluation metrics for drug discovery. To reduce the data bias and uncertain disturbance, each model is executed 10 times, and the average performance is computed. The hyperparameter selections can be found in Supplementary Material Section S15.

Data availability

All relevant data including the original network and initialization features can be downloaded from https://github.com/pengsl-lab/MSSL.git.

Code availability

The source code can be found at: https://github.com/pengsl-lab/MSSL.git. In the GitHub repository, we have provided source code that include the data processing of six SSL pretext tasks, GAT-based multi-task representation models, and MLP-based DDI or DTI predictors. Concurrently, we have added the description of how to use program. In addition, we add a license and DOI to the code. The license is GNU General Public License v3.0, and the DOI is 10.5281/zenodo.6969990.

Acknowledgement

This work was supported by NSFC Grants U19A2067, 61772543, 81973244; Science Foundation for Distinguished Young Scholars of Hunan Province (2020JJ2009); National Key R&D Program of China 2017YFB0202602, 2018YFC0910405, 2017YFC1311003, 2016YFC1302500; Science Foundation of Changsha Z202069420652, kq2004010; JZ20195242029, JH20199142034; The Funds of State Key Laboratory of Chemo/Biosensing and Chemometrics and the Cloud Brain and Major Key Project of PCL; Hunan Provincial Innovation Foundation For Postgraduate QL20210094.

Author contributions

X.W. and Y.C. conceived the original idea and developed the code for the core algorithm. S. P designed the experiment and wrote the initial version of the manuscript. F.L. and Y.Y analyzed the experimental data and edited this manuscript. Y.N.Y constructed the biomedical network data. All authors reviewed and approved the final manuscript.

Competing interests: The authors declare no competing interests.

References

[1] Dickson, M. & Gagnon, J. P. Key factors in the rising cost of new drug discovery and development. *Nat. Rev. Drug Discov.* **3**, 417-429 (2004).

[2] Scannell, J. , Blanckley, A., Boldon, H., & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191-200 (2012).

[3] Shen, W. X. et al. Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat. Mach. Intell.* **3**, 334-343 (2021).

[4] Chen, D. et al. Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nat. Commun.* **12**, 1-9 (2021).

[5] Ye, Q. et al. A unified drug-target interaction prediction framework based on knowledge graph and recommendation system. *Nat. Commun.* **12**, 1-12 (2021).

[6] Luo, Y. et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* **8**, 1-13 (2017).

[7] Cheng, F. et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun.* **9**, 1-12 (2018).
[8] Chu, Y. et al. DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Brief. Bioinformatics.* **22**, 451-462 (2021).
[9] Chu, Y. et al. DTI-MLCD: predicting drug-target interactions using multi-label learning with community detection method. *Brief. Bioinformatics.* **22**, bbaa205 (2021).
[10] Zheng, S. et al. Predicting drug-protein interaction using quasi-visual question answering system. *Nat. Mach. Intell*. **2**, 134-140 (2020).
[11] Liu, R., Wei, L., & Zhang, P. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nat. Mach. Intell.* **3**, 68-75 (2021).
[12] Ryu, J., Kim, H., & Lee, S. Deep learning improves prediction of drug-drug and drug-food interactions. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4304-E4311 (2018).
[13] Lin, S. et al. MDF-SA-DDI: predicting drug-drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. *Brief. Bioinformatics.* **23**, bbab421 (2022).
[14] Jiménez-Luna, J., Grisoni, F., & Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2**, 573-584 (2020).
[15] Zeiler, M. D., & Fergus, R. Visualizing and understanding convolutional networks. *In Proceedings of the 13th European Conference on Computer Vision* 818-833 (Springer, 2014).
[16] Hochreiter, S., & Schmidhuber, J. Long short-term memory. *Neural Comput* **9**, 1735-1780 (1997).
[17] Cheng, F., Kovács, I. A., & Barabási, A. L. Network-based prediction of drug combinations. *Nat. Commun.* **10**, 1-11 (2019).
[18] Kipf, T. N., & Welling, M. Semi-supervised classification with graph convolutional networks. *In Proceedings of the 4th International Conference on Learning Representations* (OpenReview.net, 2017).
[19] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. Graph Attention Networks. *In Proceedings of the 5th International Conference on Learning Representations* (OpenReview.net, 2018).
[20] Hamilton, W. L., Ying, R., & Leskovec, J. Inductive representation learning on large graphs. *In Proceedings of the 31st International Conference on Neural Information Processing Systems* 1025-1035 (MIT Press, 2017).
[21] Zitnik, M., Agrawal, M., & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34**, i457-i466 (2018).
[22] Ma, T., Xiao, C., Zhou, J., & Wang, F. Drug similarity integration through attentive multi-view graph auto-encoders. *In Proceedings of the 27th International Joint Conference on Artificial Intelligence* 3477-3483 (Morgan Kaufmann, 2018).
[23] Gysi, D. M. et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc. Natl. Acad. Sci. U.S.A.* **118** (2021).

[24] Wang, Z., Zhou, M., & Arnold, C. Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing. *Bioinformatics* **36**, i525-i533 (2020).
[25] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. Momentum contrast for unsupervised visual representation learning. *In Proceedings of the 33th IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9729-9738 (IEEE, 2020).
[26] Grill, J. B. et al. Bootstrap your own latent: A new approach to self-supervised learning. *In Proceedings of the 34th Conference on Neural Information Processing Systems* **33**, 21271-21284 (MIT Press, 2020).
[27] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistic* 4171-4186 (NAACL, 2019)
[28] Brown, T. B. et al. Language models are few-shot learners. *In Proceedings of the 34th Conference on Neural Information Processing Systems* **33**, 1877-1901 (MIT Press, 2020).
[29] Pham, T.H., Qiu, Y., Zeng, J., Xie, L., & Zhang, P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening. *Nat Mach Intell* **3**, 247-257 (2021).
[30] Wang, Y., Min, Y., Chen, X., & Wu, J. Multi-view Graph Contrastive Representation Learning for Drug-Drug Interaction Prediction. *In Proceedings of the 30th Web Conference* 2921-2933 (ACM, 2021).
[31] Wang, X. et al. DeepR2cov: deep representation learning on heterogeneous drug networks to discover anti-inflammatory agents for COVID-19. *Brief. Bioinformatics* **22**, 1-14 (2021).
[32] Chu, Y., et al. A transformer-based model to predict peptide-HLA class I binding and optimize mutated peptides for vaccine design. *Nat Mach Intell* **4**, 300-311 (2022).
[33] Rong, Y. et al. Self-supervised graph transformer on large-scale molecular data. *In Proceedings of the 34th Conference on Neural Information Processing Systems* **33**, 12559-12571 (MIT Press, 2020).
[34] Wang, X. et al. BioERP: biomedical heterogeneous network-based self-supervised representation learning approach for entity relationship predictions. *Bioinformatics* **37**, 4793-4800 (2021).
[35] Hu, W. et al. Strategies for pre-training graph neural networks. *In Proceedings of the 8th International Conference on Learning Representations* (OpenReview.net, 2020).
[36] Jin, W., et al. Self-supervised learning on graphs: Deep insights and new direction. Preprint at https://doi.org/10.48550/arXiv.2006.10141 (2020).
[37] Zeng, X. et al. Target identification among known drugs by deep learning from heterogeneous networks. *Chem Sci* 11, 1775-1797 (2020).
[38] Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* 9, 513-530 (2018).

[39] Deng, Y. et al. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics* **36**, 4316-4322 (2020).
[40] Belkin,M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* **15**, 1373-1396 (2003).
[41] Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V., and Smola, A. J. Distributed large-scale natural graph factorization. *In Proceedings of the 22nd International World Wide Web Conference* 37-48 (ACM, 2013).
[42] Perozzi, B., Al-Rfou, R., and Skiena, S. Deepwalk: Online learning of social representations. *In Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining* 701-710 (ACM, 2014).
[43] Liu, B., & Tsoumakas, G. Optimizing Area Under the Curve Measures via Matrix Factorization for Predicting Drug-Target Interaction with Multiple Similarities. Preprint at https://arxiv.org/abs/2105.01545 (2021).
[44] Wang, Y., Min, Y., Chen, X., & Wu, J. Multi-view graph contrastive representation learning for drug-drug interaction prediction. *In Proceedings of the 30th Web Conference* 2921-2933 (ACM, 2021).
[45] Hou, W., & Cronin, S. B. A review of surface plasmon resonance-enhanced photocatalysis. *Adv. Funct. Mater.* **23**, 1612-1619 (2013).
[46] Sigman, M., & Cecchi, G. A. Global organization of the Wordnet lexicon. *Proceedings of the National Academy of Sciences* **99**, 1742-1747 (2002).
[47] Watts, D. J., & Strogatz, S. H. Collective dynamics of 'small-world'networks. Nature **393**, 440-442 (1998).
[48] Costa, L. D. F., Rodrigues, F. A., Travieso, G., & Villas Boas, P. R. Characterization of complex networks: A survey of measurements. *Advances in physics*, **56**, 167-242 (2007).
[49] Peng, Z., Dong, Y., Luo, M., Wu, X. M., & Zheng, Q. Self-supervised graph representation learning via global context prediction. Preprint at https://arxiv.org/abs/2003.01604 (2020).
[50] Fu, G. et al. Predicting drug target interactions using meta-path-based semantic network analysis. *BMC bioinformatics* **17**, 1-10 (2016).
[51] Wu, G, Liu, J. & Yue, X. Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition. *BMC bioinformatics* **20**, 1-13 (2019).
[52] Breiman L. Random forests. *Machine learning* 45, 5-32 (2001).
[53] Cortes, C., & Vapnik, V. Support-vector networks. *Machine learning* **20**, 273-297 (1995).
[54] Yang, X., Deng, C., Dang, Z., Wei, K., & Yan, J. SelfSAGCN: Self-Supervised Semantic Alignment for Graph Convolution Network. In *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 16775-16784 (IEEE, 2021).
[55] Kapidis, G., Poppe, R., & Veltkamp, R. C. Multi-dataset, multitask learning of egocentric vision tasks. *IEEE Trans Pattern Anal Mach Intell (TPAMI)*. DOI: 10.1109/TPAMI.2021.3061479 (2021).
[56] Hernández-Lobato, D., & Hernández-Lobato, J. M. Learning feature selection

dependencies in multi-task learning. *In Proceedings of the 27th Conference on Neural Information Processing Systems.* 746-754 (MIT Press, 2013).

[57] Zhao, S., Liu, T., Zhao, S., & Wang, F. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In *Proceedings of the 33th AAAI Conference on Artificial Intelligence*. 817-824 (AAAI, 2019).

[58] Baxter, J. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning.* **28**, 7-39 (1997).

[59] Ruder, S. An overview of multi-task learning in deep neural networks. Preprint at https://doi.org/10.48550/arXiv.1706.05098 (*2017*).

[60] Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* ***4***, 279-287 (2022).

[61] Zeng, J., & Xie, P. Contrastive self-supervised learning for graph classification. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. 10824-10832 (*AAAI*, 2021).

[62] Li, T., Wang, L., & Wu, G. Self-supervision to distillation for long-tailed visual recognition. In *Proceedings of the 34th IEEE/CVF International Conference on Computer Vision*. 630-639 (IEEE, 2021).

[63] Li, Y, et al. GMSS: Graph-Based Multi-Task Self-Supervised Learning for EEG Emotion Recognition. *IEEE Trans. Affect. Comput.* DOI: 10.1109/TAFFC.2022. 3170428 (2022).

[64] Law, V., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42**, D1091-7 (2014).

[65] Hernandez, T., et al. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.* **36**, D913-8 (2008).

[66] Zhu, F., et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.* **40**, D1128-36 (2012).

[67] Keshava Prasad,T. et al. Human protein reference database 2009 update. *Nucleic Acids Res.* **37**, D767-D772 (2009).

[68] Figeys D. Mapping the human protein interactome. *Cell Res.* 18, 716-24 (2008).

[69] Oughtred, R. et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. **47**, D529-D541 (2019).

[70] Brown, A. S., & Patel, C. J. A standard database for drug repositioning. *Sci. Data*, **4**, 170029 (2017).

[71] Ursu, Oleg. et al. DrugCentral: online drug compendium. *Nucleic Acids Res*. **45**, D932-D939 (2017).

[72] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-D517 (2005)

[73] Davis, A.P. et al. The comparative toxicogenomics database: update 2013. *Nucleic Acids Res*. **41**, D1104-D1114 (2013).

[74] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. **32**, D267-D270 (2004).

[75] Watts, D. J. Networks, dynamics, and the small-world phenomenon. *Am. J. Sociol.* **105**, 493-527 (1999).
[76] Taylor, W. L. “Cloze procedure”: a new tool for measuring readability. *Journalism Quarterly* **30**, 415-433 (1953).
[77] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Model* **28**, 31-36 (1988).
[78] Vilar, S. et al. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nat. Protoc.* **9**, 2147-2163 (2014).
[79] Gotoh, O. An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705-708 (1982).
[80] Ni, P. et al. Constructing disease similarity networks based on disease module theory. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**, 906-915 (2018).
[81] Liu, P., Qiu, X., & Huang, X. Adversarial multi-task learning for text classification. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* 1-10 (ACL, 2017).
[82] Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., & Erhan, D. Domain separation networks. *In Proceedings of the 30th Conference on Neural Information Processing Systems* **29**, 343-351 (2016).
[83] Collobert, R., & Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. *In Proceedings of the 25th International Conference on Machine Learning* 160-167 (ACM, 2008).